Group Project Description

I. Introduction

This project aims to encourage you to implement and compare the three dominant methodologies in sentiment analysis of financial news sentences, i.e., 1) lexicon-based model, 2) statically learning method and 3) models based on word or sentence/doc encodings as introduced in [1] .

II. Tasks

You are required to build three predictive models with **Python** that can predict the sentiment polarity of a sentence, i.e., positive or negative. In particular, you are required to build three different models with different restrictions as following tables.

	Model-1	Model-2	Model-3
Adopted learning models	Lexicon based models	Statistical models*	Any
Provided Resources	LM Dictionary+	LM Dictionary	LM Dictionary
Extra Resources	None except for stop words	None except for stop words	Any

* Statistical models can be those models introduced in our course, including: 1) decision tree, 2) linear perceptron, 3) logistic regression, 4) SVM, 5) Naïve Bayes, 6) KNN, 7) random forest, 8) gradient boost, 9)Adaboost, 10) other ensemble models of previous base models.

+ LM Dictionary refers to the "LoughranMcDonald_SentimentWordLists_2018.xlsx" described below.

III. Provided resources

You are provided with following resources for this project:

1) **train.xlsx** contains 4338 sentences together with the sentiment labels, where 1 represent positive and -1 represent negative;
2) **train.xlsx** contains 1000 sentences with unknown sentiment label**.**
3) **test_to-submit-model-x.xlsx** provides an example to illustrate the expected format of the predicted labels for the 1000 sentences (identified by the 'rid' column) by your developed models-x (x can be 1 or 2 or 3 accordingly);

4) **LoughranMcDonald_SentimentWordLists_2018.xlsx** contains each of the LM sentiment words by category (Negative, Positive, Uncertainty, Litigious, Strong Modal, Weak Modal, Constraining). And this is the **only lexicon** that model-1 can use.
5) **References folder** contains the recommend reading for developing the three aforementioned models.

IV. Evaluation

1. **Model performance 70%**

   The performance of the three aforementioned models will be evaluated individually, and overall weight of performance evaluation for Model-1, Model-2 and Model-3 will be **20%, 30%, 20%** respectively. In particular, the performance of each model will further be evaluated from following perspectives:

   1) Novelty, practicality and rationality of proposed methodology (**10%**);
   2) Performance evaluation results measured by F1 score based on the predicted labels for the test dataset as in the submitted file (**90%**). And the raw F1 score will be normalized by the overall performance of whole class on the same model;

   For example, the F1 score of the Model-2 will account for 30%*90% = 27% of the overall performance.

2. **Individual workload and performance: 15%**
3. **Peer Evaluation: 15%**

V. Deliverables and Due Dates

1) **Predicted results (Due: May 10, 2021 at 23:59pm)**

   You are required to fill your predicted results into **3 files** based on the **3 models** developed by you, respectively. The file format is described as "test_to-submit-model-x.xlsx", and please replace the x with 1,2 or 3 to indicate which model is employed to predict the labels. For example, a file named as "test_to-submit-model-**1**.xlsx" stores the predicted labels of the sentences in "test.xlsx" by Model-**1**.

2) **Slides (Due: May 13, 2021 at 23:59pm)**

   The slides mainly introduce the main idea of your proposed approach for each built models, operations you take to enhance models performance. In addition, you also need to report the assigned job duties of each member for this project as suggested in the following table.

Member Name	Student ID	Roles and Responsibility
A	123	e.g., Model-1 (classification rules), Model-2

		(feature engineering), Model-3 (pre-training)
B	456	e.g., Model-1 (text preprocessing), Model-2 (model development), Model-3 (model tuning)

3) Presentation (Due: May 16, 2021 Morning and Afternoon)

- ✓ 10 minute presentation and 3-minute Q&A
- ✓ All team members are encouraged to be active in the presentation.

4) Project file & Peer Evaluation Form (Due: Fri. May 21, 2021 at 23:59PM)

You need to submit your Python code files together with dependent resources for the 3 developed models, respectively. You may submit this as a file attachment or a file sharing link if the sizes of file exceed the attachment limit (e.g., Baidu Yunpan).

V. Notes

- **Only electronic submission** will be required
- You need to make sure that the submitted Python code is **runnable** and can generate and save the predicted labels that are **consistent with** the labels in the files submitted by you for each model accordingly. Otherwise, the scores generated by the Python code will be adopted for grading and will incur a 30% penalty on the grade. If there is any error in your submitted Python code such that it cannot generate valid ranking scores, the whole project will be skipped for grading.
- You may consider improve your model performance in text clean and preprocess text representation and model selection and tuning in general. For example, you may try refine the stop words in general.
- You may consider the uncertainty and negation of sentences in building Model-1, try different feature representation in developing Model-2, incorporate word2vec in feature engineering, or pre-trained transformer in developing Model-3
- Do not label the sentences in "test.xlsx" manually and then use it as extra resources in developing Model-3.

References

[1] Mishev, Kostadin, et al. "Evaluation of sentiment analysis in finance: from lexicons to transformers." *IEEE Access* 8 (2020): 131662-131682.